Space-Time Earthquake Prediction: the Error Diagrams

G. MOLCHAN

International Institute of Earthquake Prediction Theory and Mathematical Geophysics, Russian Academy of Sciences, Moscow, Russia.
The Abdus Salam International Centre for Theoretical Physics, SAND Group, Trieste, Italy

E-mail: molchan@mitp.ru

Abbreviated title: The Error Diagrams

Abstract—The quality of earthquake prediction is usually characterized by a two-dimensional diagram n vs. τ, where n is the rate of failures-to-predict and τ is a characteristic of space- time alarm. Unlike the time prediction case, the quantity τ is not defined uniquely, so that the properties of the (n, τ) diagram require a theoretical analysis, which is the main goal of the present study. This note is based on a recent paper by Molchan and Keilis-Borok in GJI, 173 (2008), 1012-1017.

Key words: prediction, earthquake dynamics, statistical seismology.

1 *Introduction*

The sequence of papers (Molchan 1990, 1991, 1997, 2003) considers earthquake prediction as a decision making problem. The basic notions in this approach are the strategy, π, and the goal function, φ. Any strategy is a sequence of decisions $\pi(t)$ about an alarm of some type for a next time segment $(t, t+\delta)$, $\delta \ll 1$; $\pi(t)$ is based on the data $I(t)$ available at time t. The goal of prediction is to minimize φ, and the mathematical problem consists in describing the optimal strategy. Molchan (1997) considered the problem under the conditions in which target events form a random point process $dN(t)$ ($N(t)$ is the number of events in the interval $(0, t)$), and the aggregate $\{dN(t),\, I(t),\, \pi(t)\}$ is stationary.

Dealing with the prediction of time, Molchan (1997) considered, along with the general case, the situation in which the optimal strategy is locally optimal, i.e., is optimal for any time segment. This case arises when the goal function has the form $\varphi(n, \tau)$, where n, τ are the standard prediction characteristics/errors: n is the rate of failures-to-predict and τ the alarm time rate. The optimal strategy can then be described in much simpler terms, and can be expressed by the conditional rate of target events

$$r(t) = P\{dN(t) > 0 \,|\, I(t)\}/dt, \tag{1}$$

the loss function φ, and the error diagram $n(\tau)$. The last function can be defined as the lower bound of the set $\mathcal{E} = \{n, \tau\}$; this set consists of the (n, τ) characteristics of all the strategies based on $I(t)$. The search for the optimal strategy on a small time segment $(t, t+\delta)$ is reduced to the classical testing of two simple hypotheses such that the errors of the two kinds $(\beta(\alpha), \alpha)$ (Lehmann, 1959), converge to $(n(\tau), \tau)$ as $\delta \downarrow 0$. In statistical applications the curve $1 - \beta(\alpha)$ is known as the ROC diagram or Relative/Receiver Operating Characteristic (Swets, 1973); its limit in the case of the locally optimal strategy gives the curve $1 - n(\tau)$.

The error diagram $n(\tau)$ has proved to be so convenient a tool for the analysis of prediction methods that it began to be also used for the prediction of the space-time of target events. In that case the part of τ is played by a weighted mean of τ over space. To be specific, we divide the space G into nonintersecting parts $\{G_i\}$ and denote by τ_i the alarm time rate in G_i for the strategy π. The space-time alarm is effectively measured by

$$\tau_w = \sum_{i=1}^{k} w_i \tau_i, \quad \sum_{i=1}^{k} w_i = 1\,, \quad w_i \geq 0, \tag{2}$$

where the $\{w_i\}$ depend on the prediction goals, e.g., at the research stage of prediction one use

$$w_i = \text{area of } G_i/\text{area of } G \tag{3}$$

(Tiampo et al., 2002; Shen et al., 2007; Zechar and Jordan, 2008; Shcherbakov et al.,2008) or

$$w_i = \lambda(G_i)/\lambda(G), \tag{4}$$

where $\lambda(G)$ is the rate of target events in G (Keilis-Borok and Soloviev, 2003; Kossobokov, 2005). When dealing with the social and economic aspects of prediction, it is advisable to use weights of the form

$$w_i = \int_{G_i} p(g)\,dg \Big/ \int_{G} p(g)\,dg, \tag{5}$$

where $p(g)$ is, e.g., the density of population in G.

The $n(\tau_w)$ diagrams constructed on analogy with the error diagram are frequently ascribed also the properties of $n(\tau)$. We now mention those properties which, in the case of $n(\tau_w)$, either must be better specified or are wrong:

(a) $n(\tau)$ characterizes the limiting prediction capability of the data $\{I(t)\}$. That means that the minimum of any loss function $\varphi(n,\tau)$ with convex levels $\{\varphi \leq c\}$ is reached at the curve $n(\tau)$; (b) φ and $n(\tau)$ define the optimal strategy and its characteristics (n,τ); (c) the diagonal D of the square $[0,1]^2$, $n+\tau=1$, is the antipode of $n(\tau)$, because it describes the characteristics of *all* trivial strategies which are equivalent to random guess strategies. Therefore, the maximum distance between $n(\tau)$ and D, i.e., $\max_\tau(1-n(\tau)-\tau)/\sqrt{2}$, characterizes the prediction potential of $\{I(t)\}$; (d) $1-n(\tau)$ is a ROC diagram arising in the testing of simple statistical hypotheses.

Molchan and Keilis-Borok (2008) recently considered the prediction of the space-time of target events under conditions where the optimal strategies coincide with the locally optimal ones (the word "locally" now also refers to both space and time). This paper gives a correct extension of the error diagram, which provides the key to the understanding of the information contained in an $n(\tau_w)$ diagram. The present note supplements the abovementioned study. We refine the structure of the error diagram for space-time prediction and analyze the properties of two-dimensional $n(\tau_w)$ diagrams.

2 *The Error Diagram*

We quote the main result by Molchan and Keilis-Borok (2008) relevant to the prediction of space-time for target events.

Let $\{G_i\}$ be some partition of G into nonintersecting regions. The prediction of location means the indication of $\{G_i\}$ where a target event will occur. Consequently, the model of target events in G is the stationary random vector point process

$$d\mathbf{N}(t) = \{dN_1(t), \dots, dN_k(t)\}$$

whose components describe target events in $\{G_i\}$. We shall consider the binary yes/no prediction with the decisions

$$\boldsymbol{\pi}(t) = \{\pi_1(t), \ldots, \pi_k(t)\}, \quad t = n\delta$$

of the form

$$\pi_i(t) = \begin{cases} \text{alarm in } G_i \times (t, t+\delta) \\ \text{no alarm in } G_i \times (t, t+\delta) \end{cases}$$

The decision $\boldsymbol{\pi}(t)$ is based on the data $I(t)$ that are available at time t.

Under certain conditions, namely, the aggregate $\{d\mathbf{N}(t),\, I(t),\, \boldsymbol{\pi}(t)\}$ is ergodic and stationary, and moreover $P\{\sum_{i=1}^k dN_i(t) > 1\} = o(dt)$, the basic characteristics of the strategy $\pi = \{\boldsymbol{\pi}(t)\}$ are defined as the limit of its empirical means. We have in view the rate of failures-to-predict n and the vector

$$\boldsymbol{\tau} = (\tau_1, \ldots, \tau_k),$$

which determines the alarm time rate in the $\{G_i\}$. . The quantities $(n, \boldsymbol{\tau})$ are defined for any small δ. We shall assume that n and $\boldsymbol{\tau}$ have limits as $\delta \downarrow 0$, for which we retain the same notation. The passage to the limit is not a restriction, since the data may reflect the seismic situation with a fixed time delay.

The set of $(n, \boldsymbol{\tau})$ characteristics for different strategies π based on $\{I(t)\} = I$ is a convex subset in the $(k+1)$-dimensional unit cube, i.e., the error set

$$\mathcal{E}(I) = \{(n, \boldsymbol{\tau})_\pi \,:\, \pi \quad \text{based on} \quad I\} \subseteq [0,1]^{k+1}, \tag{6}$$

(see Fig. 1). The set $\mathcal{E}$ contains the simplex

$$\mathbf{D} = \{(n, \boldsymbol{\tau}) \,:\, n + \sum_{i=1}^{k} \lambda_i \tau_i / \lambda = 1,\, 0 \le n, \tau_i \le 1\}, \tag{7}$$

where $\lambda_i = \lambda(G_i)$. The set (7) describes strategies that are equivalent to the random guess strategies. For indeed, if an alarm is declared in G_i with the rate τ_i, then $\lambda_i \tau_i / \lambda$ will give the rate of random successes in G_i. The equality in (7) i.e., $1 - n = \sum_{i=1}^{k} \lambda_i \tau_i / \lambda$, means that the success rate is identical with the rate of random successes. Such strategies will be called *trivial.*

The boundary of $\mathcal{E}$, viz., $n(\boldsymbol{\tau})$, which lies below the hyperplane (7), will be called the *error diagram.* To describe the properties of $n(\boldsymbol{\tau})$, we define the loss function φ. This will be a function of the form $\varphi(n, \boldsymbol{\tau})$ that is nondecreasing in each argument and for which any level set, $\{\varphi \leq c\}$, is convex.

The following is true.

2.1. The minimum of $\varphi(n, \boldsymbol{\tau})$ on $\mathcal{E}$ is reached on the surface $n(\boldsymbol{\tau})$. The point of the minimum, Q, is found as the point where the suitable level $\{\varphi \leq c\}$ is tangent to $n(\boldsymbol{\tau})$ (see Fig. 1). The coordinates of $Q = (n, \boldsymbol{\tau})$ define the characteristics of the optimal strategy with respect to the goal function φ;

2.2. The optimal strategy declares an alarm in $G_i \times (t, t + \delta)$, $\delta \ll 1$ as soon as

$$r_i(t) = P\{\delta N_i(t) > 0 \mid I(t)\} / \delta \geq r_{0i} \tag{8}$$

and declares no alarm otherwise;

2.3. The threshold r_{0i} depends on φ, e.g., if

$$\varphi = a\lambda n + \sum_{i=1}^{k} b_i \tau_i \tag{9}$$

then $r_{0i} = b_i / a$. In the general case one has

$$r_{0i} = -\lambda \frac{\partial \varphi}{\partial \tau_i} \Big/ \frac{\partial \varphi}{\partial n}(Q).$$

The result described above yields an important corollary:

2.4. The error diagram for space-time prediction in $G = \{G_i\}$ based on $\{I(t)\}$ admits of the representation

$$n(\tau_1, \dots, \tau_k) = \sum_{i=1}^{k} \lambda_i n_i(\tau_i)/\lambda, \tag{10}$$

where $n_i(\tau)$ is the error diagram for time prediction in G_i based on the same data $\{I(t)\}$.

Proof. Consider such a loss function (9) that the hyperplane $\varphi = c$ is tangent to $n(\boldsymbol{\tau})$ at $\boldsymbol{\tau}_0 = (\tau_{01}, \dots, \tau_{0k})$. The optimal strategy thus has the form (8) with $r_{0i} = b_i/a$ and the errors $(n(\boldsymbol{\tau}_0), \boldsymbol{\tau}_0)$. However, the strategy for time prediction in G_i of the form (8) minimizes the loss function $\varphi_i = a\lambda_i n + b\tau$ (Molchan, 1997). The point of the minimum has the coordinate $\tau = \tau_{0i}$, hence the other coordinate is $n = n_i(\tau_{0i})$. Consequently, the collective strategy (8) minimizes

$$\sum_{i=1}^{k} \varphi_i = a\lambda\left(\sum_{i=1}^{k} \lambda_i n_i/\lambda\right) + \sum_{i=1}^{k} b_i \tau_i \tag{11}$$

and has $n = \sum_{i=1}^{k} \lambda_i n_i(\tau_{0i})/\lambda$ as the rate of failures-to-predict. The right-hand side of (11) is identical with $\varphi(n, \boldsymbol{\tau})$. It follows that (10) is true with $n = n(\boldsymbol{\tau}_0)$, since the strategy (8) also minimizes (9). Since $\boldsymbol{\tau}_0$ is arbitrary, the corollary is proven.

3 *The reduced error diagrams*

Usually regional error diagrams $n_i(\tau)$ are poorly estimated, so that for practical purposes the result of a space-time prediction is represented by the two-dimensional diagram $n(\tau_w)$, $\tau_w = \sum_{i=1}^{k} w_i \tau_i$ where the weights are $w_i \geq 0$ and $\sum_{i=1}^{k} w_i = 1$. This is obtained from the set of "errors" $\mathcal{E}_w = \{(n, \tau_w)\}$ as its lower boundary.

Relation (10) can be used to analyze the properties of $n(\tau_w)$ diagrams. Later we shall use the following notation: if the set B is the image of $A = \{(n, \boldsymbol{\tau})\}$ by the mapping

$$\gamma_w \,:\, (n, \boldsymbol{\tau}) \to (n, \tau_w), \quad \tau_w = \sum_{i=1}^{k} w_i \tau_i \,,$$

then $B = A_w$; in particular, the image of $\boldsymbol{\tau}$ is τ_w, the image of $\mathcal{E}$ is $\mathcal{E}_w$, while the image of D (see (7)) is D_w.

The following is true.

3.1. $\mathcal{E}_w$ is a convex subset of the square $[0,1]^2$ that contains the diagonal $\tilde{D} \,:\, n + \tau_w = 1$;

3.2. D_w is a convex subset of $\mathcal{E}_w$; D_w degenerates to the diagonal of the unit square, if and only if $w_i = \lambda_i / \lambda$, $i = 1, \ldots, k$;

3.3. D_w can be obtained as the convex hull of points of the form

$$n = 1 - \sum_{i=1}^{k} \lambda_i \varepsilon_i \,, \quad \tau_w = \sum_{i=1}^{k} w_i \varepsilon_i, \tag{12}$$

where $\{\varepsilon_i\}$ are all possible sequences of 0 and 1 (see Fig. 2).

In particular, let $w_1 = \ldots = w_k$ (this will be the case for (3) when G is divided into equal parts). Then the convex minorant of the (n, τ_w) points:

$$(1,0),\, (1 - \lambda_{(k)}, 1/k), \ldots, \left(1 - \sum_{i=1}^{p} \lambda_{(k-i+1)}, p/k\right), \ldots, (0,1)$$

gives the lower boundary of D_w, while the concave majorant of the points

$$(1,0),\, (1 - \lambda_{(1)}, 1/k), \ldots, \left(1 - \sum_{i=1}^{p} \lambda_{(i)}, p/k\right), \ldots, (0,1)$$

gives the upper boundary of D_w. Here, $\lambda_{(1)} \leq \ldots \leq \lambda_{(k)}$ are the $\{\lambda_i\}$ arranged in increasing order.

3.4. Except for trivial cases, the image of the error diagram $n(\boldsymbol{\tau})$ is a two-dimensional set (see Fig. 2) with the lower boundary $n(\tau_w)$ and the upper boundary $n^+(\tau_w)$. In the regular case, i.e., $\varphi_i(0) = 1$, $i = 1, \ldots, k$, one has

$$n^+(x) = \max_{i,\boldsymbol{\varepsilon}}\{\lambda_i/\lambda \cdot n_i(x/w_i - a_i(\boldsymbol{\varepsilon})) + b_i(\boldsymbol{\varepsilon})\}, \tag{13}$$

where

$$\begin{aligned}
\boldsymbol{\varepsilon} &= (\varepsilon_1, \ldots, \varepsilon_k), \varepsilon_i = 0, 1, \\
a_i(\boldsymbol{\varepsilon}) &= \sum_{j \neq i} w_j \varepsilon_j / w_i\,, \\
b_i(\boldsymbol{\varepsilon}) &= \sum_{j \neq i} \lambda_j (1 - \varepsilon_j)/\lambda,
\end{aligned}$$

and the maximum is taken over such i and (0,1) sequences $\boldsymbol{\varepsilon}$, for which the argument of n_i in (13) makes sense, i.e., is in $[0, 1]$.

If $\{n_i(\tau)\}$ are piecewise smooth and $n_i(0) = 1$, $i = 1, \ldots, k$, then the image of $n(\boldsymbol{\tau})$ degenerates to a one-dimensional curve, if and only if $\{I(t)\}$ is trivial, i.e., $1 - n(\boldsymbol{\tau}) = \sum_{i=1}^k \lambda_i \tau_i / \lambda$ and $w_i = \lambda_i / \lambda$, $i = 1, \ldots, k$.

3.5. The curve $n(\tau_w)$ represents those strategies which are optimal for loss functions of the form

$$\varphi(n, \boldsymbol{\tau}) = \psi(n, \tau_w), \quad \tau_w = \sum_{i=1}^k w_i \tau_i. \tag{14}$$

To be specific, if $(n, \boldsymbol{\tau}) = Q$ are the optimal prediction characteristics with respect to the goal function of the form (14), then Q_w belongs to the $n(\tau_w)$ diagram. In addition, Q_w is the point at which the curve $n(\tau_w)$ is tangent to the suitable level set of ψ.

3.6. The strategy that optimizes (14) declares an alarm in $G_i \times (t, t+\delta)$ as soon as

$$r_i(t)/w_i \geq c, \tag{15}$$

where the threshold c is independent of G_i and r_i is given by (8). According to **2.3**,

$$c = \lambda \frac{\partial \psi}{\partial \tau_w} \Big/ \frac{\partial \psi}{\partial n}(Q_w).$$

In particular, if $\varphi = an + b\sum_{i=1}^{k} w_i \tau_i$, then $c = \lambda b/a$. If $w_i = \lambda_i/\lambda$, then (15) will have the form $r_i(t)/\lambda_i \geq c\lambda$, where the left-hand side is known as the probability gain.

3.7. For any point Q in the error diagram we can find such weights $\{w_i\}$ that Q_w will lie in the reduced (n, τ_w) diagram, i.e., any optimal strategy can be represented by a suitable (n, τ_w) diagram . The desired weights are

$$w_i = -\frac{\partial n}{\partial \tau_i}(Q)/c\,,$$

where c is a normalizing constant. The point Q determines the optimal prediction characteristics with respect to the loss function

$$\varphi = n + c\sum_{i=1}^{k} w_i \tau_i.$$

3.8. The curve $1 - n(\tau_w)$ can be interpreted as a ROC diagram if and only if $w_i = \lambda_i/\lambda$, $i = 1, \ldots, k$.

The ROC property of a (n, τ_w) diagram means that we can treat (n, τ_w) characteristics as errors of the two kinds (β, α) in hypothesis testing: H_1 vs. H_0, i.e.,

$$\beta = P(H_0 \,|\, H_1) = n, \quad \text{and} \quad \alpha = P(H_1 \,|\, H_0) = \tau_w \tag{16}$$

and $\alpha + \beta = 1$, if the prediction data $\{I(t)\}$ are trivial.

In the case $w_i = \lambda_i/\lambda$ the measures $P(\cdot \,|\, H_j)$, $j = 0, 1$ can be specified as follows. Both measures define probabilities for events $\omega = \{I(t), \nu = i\}$, where ν is the random index of a subregion and has the distribution $P(\nu = i) = \lambda_i/\lambda := p_i$. The measure related to the H_0 hypothesis is

$$P(d\omega \,|\, H_0) = P_0(dI)p_i, \quad \nu(\omega) = i, \tag{17}$$

where P_0 is the stationary measure on $I(t)$ induced by the process $\{dN(t),\, I(t),\, \pi(t)\}$. In the H_1 case

$$P(d\omega \,|\, H_1) = r_i(t)/\lambda_i \cdot P(d\omega \,|\, H_0), \quad \nu(\omega) = i, \tag{18}$$

where $r_i(t)$ is given by (8).

It is better to say that testing H_1 vs. H_0 for the case $G = \{G_i\}$ involves two points: a random choice of G_i with probabilities $p_i = \lambda_i/\lambda$, $i = 1, \dots, k$ and testing H_1 vs. H_0 for the relevant subregion. The second point is considered in (Molchan and Keilis-Borok, 2008).

The following is a nontrivial corollary of the previous statement:

3.9. For the regular case, $n_i(0) = 1$, $i = 1, \dots, k$ and $\{w_i\} = \{\lambda_i/\lambda\}$, one has

$$\int_0^1 f\left(-\frac{dn_\lambda}{d\tau}\right) d\tau = \sum_{i=1}^k p_i \int_0^1 f\left(-\frac{dn_i}{d\tau}\right) d\tau, \quad p_i = \lambda_i/\lambda \tag{19}$$

where f is any continuous function and $n_\lambda(\tau)$ is an alternative notation for the $n(\tau_w)$ diagram in the special case $w_i = \lambda_i/\lambda$, $i = 1, \dots, k$.

If $f = x \log x$, the quantity

$$I_i = \int_0^1 f\Big(-\frac{dn_i}{d\tau}\Big)\, d\tau = \int_0^1 \ln\Big(-\frac{dn_i}{d\tau}\Big)\, dn_i \tag{20}$$

is known in time prediction as the *Information score* (see Kagan, 2007 and Harte & Vere-Jones, 2005).

Comments. In the non-regular case, $n_\lambda(0) < 1$, the score (19) is equal to ∞ for unbounded $f(x)$ at $x = \infty$, e.g., $f = x \log x$. Therefore the scores (19), (20) are unstable. (Extensive literature on skill scores can be found in Jolliffe & Stephenson, 2003; see also Molchan, 1997 and Harte & Vere-Jones,2005). Here we mention only the *area skill score* which is used as a stable score (Zechar & Jordan,2008). A linear transformation of this score looks as follows:

$$A = 2\int_0^1 (1 - n_\lambda(\tau) - \tau)\, d\tau, \quad 0 \le A \le 1. \tag{21}$$

Due to convexity of $n_\lambda(\tau)$ the area under the integrand is approximated by a triangle from within and by the trapezium from the outside. Therefore

$$H \le A \le H(2 - H),$$

where

$$H = \max_\tau (1 - n_\lambda(\tau) - \tau), \quad 0 \le H \le 1.$$

Thus $\widehat{A} = H(3 - H)/2$ is a good estimate of A, because

$$|A - \widehat{A}| \le H(1 - H)/2 \le 1/8. \tag{22}$$

The empirical estimate of the H skill score is unstable for a small number of target events. Due to (22) the same holds for the area skill score.

The H score is convenient for statistical analysis because its empirical estimate is identical in distribution with the Kolmogorov-Smirnov statistics D_N^+ (Bolshev & Smirnov, 1983), provided $\sum N_i(T) = N$ and $\{dN_i\}$ are independent and Poissonian.

4 *Proof*

We are going to prove the statements **3.1 - 3.9**.

Proof for **3.1, 3.2**. Obviously, the projection γ_w preserves the property of convexity. Therefore, $\mathcal{E}_w$ and D_w are convex at the same time as are $\mathcal{E}$ and D. If D_w degenerates to the diagonal $\tilde{D} \, : \, n + \tau_w = 1$, then the simplex D is given by any of the two equations: $n + \sum_{i=1}^k w_i \tau_i = 1$ and $n + \sum_{i=1}^k \lambda_i \tau_i / \lambda = 1$. Hence $w_i = \lambda_i / \lambda$.

Proof of **3.3.** The simplex D is the convex hull of $(n, \boldsymbol{\tau})$ points of the form $Q(\varepsilon) = (1 - \sum \lambda_i \varepsilon_i / \lambda, \, \varepsilon_1, \ldots, \varepsilon_k)$, where $\varepsilon_i = 0, 1$. Accordingly, D_w is the convex hull of the $Q_w(\varepsilon)$, see (12).

Proof of **3.4.** This statement follows intuitively from dimensionality considerations: the k-dimensional surface $n(\boldsymbol{\tau})$ with $k > 1$ is projected onto the (n, τ_w) plane, hence its image cannot be single-dimensional in the generic case.

In order to prove (13), we note that a convex function on the simplex $S_n = \{\sum_{i=1}^k \tau_i w_i = u, \, 0 \le \tau_i \le 1\}$ reaches its maximum at one of the edges, specifically, at a point of the form

$$\boldsymbol{\tau} = (\varepsilon_1, \ldots, \varepsilon_{i-1}, x, \varepsilon_{i+1}, \ldots, \varepsilon_k), \quad \varepsilon_j = 0; 1.$$

The use of (10) gives (13).

Suppose the upper and lower boundaries of the image of $n(\boldsymbol{\tau})$ are identical and the $\{n_i(\tau)\}$ are piecewise smooth functions. Consider all $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_k)$

for which

$$\sum_{i=1}^{k} \lambda_i n(\tau_i)/\lambda = n_0, \quad \sum_{i=1}^{k} w_i \tau_i = \tau_w, \quad n_0 = n(\tau_w),$$

where τ_w is fixed.

Varying, e.g., τ_1 and τ_2, we have after differentiation:

$$\lambda_1 n_1'(\tau_1)\tau_1' + \lambda_2 n_2'(\tau_2) = 0, \quad \tau_1' = -w_2/w_1. \tag{23}$$

If τ_1, τ_2 are points of smoothness of $n_i(\tau)$, $i = 1, 2$, then repeated differentiation of (23) will give

$$\lambda_1 n_1''(\tau_1)(w_2/w_1)^2 + \lambda_2 n_2''(\tau_2) = 0.$$

However, $n_i''(\tau_i) \geq 0, \;\; i = 1, 2.$. Hence $n_i''(\tau_i) = 0$, i.e., $n_i(\tau)$ are locally linear at all points of smoothness. Since $n_i(\tau)$ are piecewise smooth, it follows that for any discontinuous point τ_1 of $n_1(\cdot)$ one can find a point τ_2 where $n_2(\cdot)$ will be smooth. Consequently, when n_1 is discontinuous at τ, one should replace $n_1'(\tau_1)$ with $n_1'(\tau_1 + 0)$ and $n_1'(\tau_1 - 0)$ in equation (23). But then we have from (23) that $n_1'(\tau)$ is continuous at τ_1; hence all functions $n_i(\tau)$ are linear. Taking the boundary conditions $n_i(0) = 1$ and $n_i(1) = 0$ into account, we have $n_i(\tau) = 1 - \tau$. However, in that case one has $\mathcal{E} = D$, and, in virtue of **3.2**, $w_i = \lambda_i/\lambda$.

Proof of **3.5.** Let Q_w be the point where the convex set $\{\psi \leq c\}$ is tangent to the convex curve $n(\tau_w)$. The function ψ reaches its minimum at the point Q_w on $\mathcal{E}_w$, because the sets $\{\psi \leq c\}$ are increasing with increasing c. Since $Q_w \in \mathcal{E}_w$, the preimage $Q = (n, \boldsymbol{\tau}) \in \mathcal{E}$. At this point $\varphi(Q) = \psi(Q_w)$ reaches its minimum on $\mathcal{E}$, hence Q belongs to the surface $n(\tau)$.

Proof of **3.6.** follows from **2.3.**

Proof of **3.7.** Let $Q = (n_0, \tau_{01}, \ldots, \tau_{0k})$ belong to $n(\boldsymbol{\tau})$. If $w_i = -\frac{\partial n}{\partial \tau_i}(Q)/c$, then the equation

$$n + c\sum_{i=1}^{k} w_i\tau_i = n_0 + c\sum_{i=1}^{k} w_i\tau_{i0} \tag{24}$$

defines the tangent plane to $n(\boldsymbol{\tau})$. Since $n(\boldsymbol{\tau})$ is convex and decreasing, it follows that $w_i \geq 0$ and $\mathcal{E}$ lie on the same side of the plane (24). Consequently, a strategy having the characteristics $Q = (n_0, \tau_{01}, \ldots, \tau_{0k})$ optimizes the losses $\varphi = n + c\sum_{i=1}^{k} w_i\tau_i$. Using **3.5**, we complete the proof.

Proof of **3.8.** By (10) and (16) one has

$$\beta = n = \sum_{i=1}^{k} \lambda_i/\lambda \cdot n_i(\tau_i), \quad \alpha = \tau_w = \sum_{i=1}^{k} w_i\tau_i.$$

In the trivial case of $I(t)$, one has $n_i(\tau) = 1 - \tau$ and $\alpha + \beta = 1$. Hence

$$\beta = 1 - \sum_{i=1}^{k} \lambda_i/\lambda \cdot \tau_i, \quad \alpha = \sum_{i=1}^{k} w_i\tau_i = 1 - \beta,$$

i.e., $w_i = \lambda_i/\lambda$, $i = 1, \ldots, k$.

Suppose that $\{w_i\} = \{\lambda_i/\lambda\}$. The likelihood ratio of measures (17) and (18) at the point $\omega = (J(t), j)$ is

$$L(\omega) = P(d\omega \mid H_1)/P(d\omega \mid H_0) = r_j(t)/\lambda_j.$$

Accepting the hypothesis H_1 as soon as $L(\omega) > c$ and H_0 otherwise, one has

$$\begin{aligned}
\alpha &= \int_{L>c} P(d\omega \mid H_0) = \sum_{j=1}^{k} E\mathbf{1}_{(r_j/\lambda_j>c)} \cdot \lambda_j/\lambda = \sum_{j=1}^{k} \tau_j\lambda_j/\lambda = \tau_w, \\
\beta &= \int_{L>c} L(w)P(d\omega \mid H_0) = \sum_{j=1}^{k} Er_j/\lambda_j \cdot \mathbf{1}_{(r_j/\lambda_j<c)} \cdot \lambda_j/\lambda = \sum_{j=1}^{k} n_j(\tau_j)\lambda_j/\lambda = n.
\end{aligned}$$

Here we have used **2.1.** and **2.2.**

Proof of **3.9.** Let us consider a testing problem: H_1 vs. H_0 with the errors $\beta = P_1(L < c)$ and $\alpha = P_0(L \geq c)$ where $L(\omega) = dP_1/dP_0$ is the likelihood ratio. Obviously

$$E_0 f(L) := \int f(L(\omega))dP_0(\omega) = \int f(c)dF_L(c),$$

where F_L is the distribution of L with respect to the measure P_0. But $d\beta = cdF(c)$ and $d\alpha = -dF(c)$. Therefore

$$E_0 f(L) = \int_0^1 f\biggl(-\frac{d\beta}{d\alpha}\biggr)\, d\alpha.$$

Applying this relation to the case (16), (17), (18), one has

$$\begin{aligned} \int_0^1 f\biggl(-\frac{dn_\lambda}{d\tau}\biggr)\, d\tau &= E_0 f(L) = \sum_{i=1}^k Ef\biggl(\frac{r_i(t)}{\lambda_i}\biggr)p_i = \\ = \sum_{i=1}^k Ef(L_i)p_i &= \sum_{i=1}^k p_i \int_0^1 f\biggl(-\frac{dn_i}{d\tau}\biggr)\, d\tau \end{aligned}$$

Here L_i is the likelihood ratio dP_1/dP_0 for G_i.

5 *Conclusion and Discussion*

1.*Results.* In the case of time prediction, the error set $\mathcal{E}$ is organized as follows: all trivial strategies concentrate on the diagonal $n + \tau = 1$ of the square $[0, 1]^2$, while the optimal strategies are on the lower boundary of $\mathcal{E}$, viz. $n(\tau)$. In the case of time-space prediction, the two-dimensional images of $\mathcal{E}$, i.e., $\mathcal{E}_w$, are organized differently: the diagonal $n + \tau_w = 1$ does not include all trivial strategies, and the (n, τ_w) diagram does not include all optimal strategies (see Fig. 2).

Nevertheless, $n(\tau_w)$ is a convenient tool to visualize such optimal strategies as are suitable for a trade-off between n and τ_w. However, if $\{w_i\} \neq \{\lambda_i/\lambda\}$, then the distance of $n(\tau_w)$ from the diagonal $n + \tau_w = 1$ does not tell us anything about the prediction potential of the relevant strategies. To learn something about this potential, we need the image of trivial strategies D on the (n, τ_w) plane. The lower boundary of D_w may be very close to the ideal strategy with the errors (0, 0).

Let us consider an example. The relative intensity (RI) method (Tiampo et al., 2002) predicts the target event in that location where the historical seismicity rate, $f(g)$, is the highest, $f > c$. The RI is a typical example of a trivial strategy occasionally employed as an alternative to meaningful prediction techniques (see, e.g., Marzocchi et al., 2003). By the RI method, $\tau_i = 1$, if $f > c$ in the *i-th* bin and $\tau_i = 0$ otherwise. If $\{w_i = \lambda_i/\lambda\}$, then

$$1 - n = \int_{f>c} f(g)\, dg = \tau_w,$$

i.e., $n + \tau_w = 1$ for any level c. If $w_i = |G_i|/|G|$, where $|G|$ is the area of G, then the curve $n(\tau_w)$ can be obtained by using (12) (see also Zechar and Jordan,2008). The curve passes close to (0,0), if most of the target events occur in a relatively small area, say, λ_1/λ is close to 1 and w_1 is close to 0.

One gets a unique set of weights by choosing $w_i = \lambda_i/\lambda$ (see **3.2**, **3.8**). It is only in this particular case that all trivial strategies are projected onto the diagonal $\tilde{D}\, :\, n + \tau_w = 1$, and $1 - n(\tau_w)$ is a ROC curve. Besides, the projection on the (n, τ_w) plane preserves the relative distance between any strategy and the set of trivial strategies . To be more specific, the following relations are true:

$$1 - n - \sum_{i=1}^{k} \tau_i \lambda_i/\lambda = \frac{\rho(Q, D)}{\rho(O, D)} = \frac{\rho(Q_w, \tilde{D})}{\rho(O_w, \tilde{D})} = 1 - n - \tau_w \qquad (25)$$

(Molchan and Keilis-Borok, 2008). Here, ρ is the Euclidean distance, e.g., $\rho(O, D)$ is the distance from $Q = (n, \boldsymbol{\tau})$ to the hyperplane $D = \{n + \sum \tau_i \lambda_i / \lambda = 1\}$, and $O = (0, 0 \dots 0)$ corresponds to the ideal strategy. The right-hand side of (25) is known in the contingency table analysis as the HK skill score (Hanssen-Kuiper, 1965). Consequently, when $\{w_i = \lambda_i/\lambda\}$, the quantity $H = \max\limits_{\tau_w}(1 - n(\tau_w) - \tau_w)$ gives the greatest relative distance between the optimal and the trivial strategies.

The choice of $\{w_i\}$ at the research stage instead of $\{\lambda_i/\lambda\}$ is justified by difficulties in the way of estimating the $\{\lambda_i\}$. This justification is illusory, however. One must know the lower boundary of D_w in order to answer the question of how nontrivial the $n(\tau_w)$ diagram is. But this again requires knowledge of the $\{\lambda_i\}$ (see (12) and Fig. 2).

2. *The relation to the SDT*. In recent years the studies in earthquake prediction are actively using the Signal Detection Theory (SDT) developed in the late 1980s in the atmospheric sciences (see, e.g., Jalliffe and Stephenson, 2003 and the references therein).The main object of this theory is a warning system, which characterizes the state of hazard by a scalar quantity ξ. The system is tested by making $K \gg 1$ trials in which the *i-th* event $\{\xi > u\}$ is interpreted as an alarm, $\widehat{x}_i = 1$, otherwise $\widehat{x}_i = 0$. The results are compared with observations $x =$ Yes or No with respect to a target event. Any dependence between the members of the sequence $\{(\widehat{x}_i, x_i)\}$ is ignored a priori. It is required only that the rate of target events ($x =$ Yes) should be $0 < s < 1$. This condition is essential for getting an acceptable estimate for the simultaneous distribution of $(\widehat{x}_i, x_i)$. Note that $s = 0$ in our approach.

Two problems are formulated: assessing the prediction performance and choosing the threshold u in a rational manner. The first problem is attacked using the 2×2 contingency table of forecasts and the second by using the ROC diagram related to the hypothesis testing about the conditional distribution of ξ given $x =$ Yes and given $x =$ No.

In our terminology this situation is one with discrete "time" where the data I in a trial are given by ξ. Therefore, the SDT is equivalent to the analysis of the time prediction of earthquakes using a specified precursor/algorithm, even though the prediction of large earthquakes involves $s \ll 1$. The ROC/$n(\tau)$ diagram then quantifies the predictive potential of a precursor, ξ in this case. All meaningful strategies are functions of ξ, hence reduce to choosing the level u.

In the case of any data, $I(t)$, $n(\tau)$ characterizes the prediction performance of $\{I(t)\}$ and gives the lower bound to ROC curves for any algorithm based on $\{I(t)\}$. The studies of Molchan (1990, 1997) answer the question of how the quantity ξ should be constructed for the original prediction data and why the relation to hypothesis testing arises at all.

The gist of the matter lies in the fact that the 2×2 contingency table is defined by three parameters (n, τ, s), and the program of prediction optimization is formulated, explicitly or implicitly, in terms of that table. As a result, we have to deal with local optimal strategies only. When real time is incorporated in the SDT framework, there arise additional parameters that are important for seismological practice, e.g., the rate of connected alarms (alarm clusters) ν. The optimization of the loss function $\varphi = an + b\tau + c\nu$ at once gets us beyond the SDT framework and its tools. The strategies that optimize φ are not locally optimal, and can be found from Bellman-type equations (Molchan and Kagan, 1992; Molchan, 1997).

The use of the SDT approach in space-time prediction imposes a rather unrealistic limitation: the spatial rate of target events must be homogeneous. Otherwise, the ROC diagram looses its meaning and becomes a (n, τ_w) diagram (see Fig. 2).

Acknowledgements

This work was supported by the Russian Foundation for Basic Research through grant 08-05-00215. I thank D.L. Turcotte for useful discussions, which have stimulated the writing of the present paper.

References

Bolshev, L.N. and Smirnov, V.N., Tables of mathematical statistics, (Nauka, Moscow 1983).

Harte, D. and Vere-Jones, D. (2005), The Entropy Score and Its Uses in Earthquake Forecasting, Pure Appl. Geophys. 162, 1229-1253.

Hanssen, A.W. and Kuipers, W.J.A. (1965), On the relationship between the frequency of rain and various meteorological parameters. Modedeelingen en Verhandelingen, Royal Notherlands Meteorological Institute, 81

Jolliffe, I.T. and Stephenson, D.B. (eds.), Forecast Verification: a Practitioner's Guide in Atmospheric Science (John Wiley & Sons, Hoboken 2003).

Kagan, Y.Y. (2007), On Earthquake Predictability Measurement: Information Score and Error Diagram, Pure Appl. Geophys. 164, 1947- 1962.

Keilis-Borok, V.I. and Soloviev, A.A. (eds.), Nonlinear Dynamics of the Lithosphere and Earthquake Prediction (Springer-Verlag, Berlin-Heidelberg 2003).

Keilis-Borok, V.I., Shebalin, P., Gabrielov, A., Turcotte, D. (2004). Reverse Tracing of Short-term Earthquake Precursors, Phys. Earth. Planet. Inter. 145, 75-85.

Kossobokov, V.G. (2005), Earthquake Prediction: Principles, Implementation, Perspectives, Computational Seismology, Iss. 36-1, 3-175, (GEOS, Moscow).

Lehmann, E.L., Testing Statistical Hypotheses (J. Wiley& Sons. New York 1959).

Marzocchi, W., Sandri, L., and Boschi, E.(2003), On the Validation of Earthquake-forecasting Models: the Case of Pattern Recognition Algorithms, Bull. Seism. Soc. Am. 93, 5, 1994-2004.

Molchan, G.M. (1990), Strategies in strong earthquake prediction, Phys. Earth Planet. Inter. 61(1-2), 84-98

Molchan, G.M. (1991), Structure of Optimal Strategies of Earthquake Prediction Tectonophysics 193, 267-276.

Molchan, G.M. (1997), Earthquake Prediction as a Decision Making Problem, Pure Appl. Geophys. 149, 233-247.

Molchan, G.M., Earthquake Prediction Strategies: a Theoretical Analysis. In Nonlinear dynamics of the Lithosphere and Earthquake Prediction (eds. Keilis-Borok, V.I. and Soloviev, A.A.) (Springer-Verlag, Berlin-Heidelberg 2003), pp.209-237.

Molchan, G.M. and Kagan, Y.Y. (1992), Earthquake Prediction and its Optimization, J. Geophys. Res. 97, 4823-4838.

Molchan, G.M. and Keilis-Borok, V.I., (2008), Earthquake Prediction: Probabilistic Aspect, Geophys. J. Int. 173, 1012-1017.

Shcherbakov, R., Turcotte, D.L., Holliday, J.R., Tiampo, K.F., and Rundle, J.B. (2008), A Method for Forecasting the Locations of Future Large Earthquakes: An Analysis and Verification, Geophys. Res. Lett., DOI: 1010291 (in press).

Shen, Z.-K., Jackson, D.D., and Kagan, Y.Y. (2007), Implications of Geodetic Strain Rate for Future Earthquakes, with a Five-Year Forecast of M5 Earthquakes in Southern California, Seismol. Res. Lett. 78(1), 116-120.

Swets, J.A. (1973), The Relative Operating Characteristic in Psychology, Science 182, 4116, 990-1000.

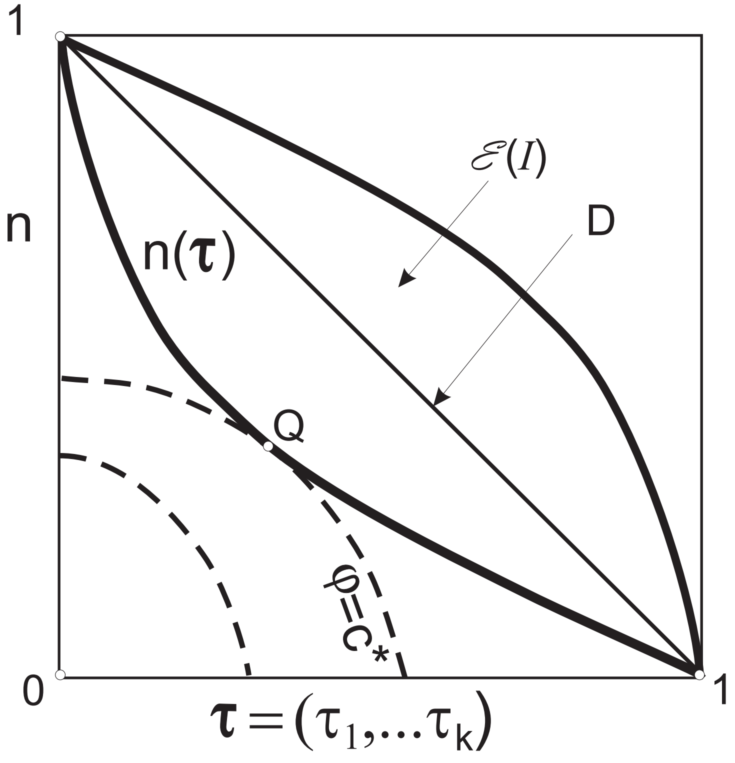
Tiampo, K.F., Rundle, J.B., McGinnis, S., Gross, S., and Klein, W. (2002), Mean Field Threshold Systems and Phase Dynamics: An Application to Earthquake Fault Systems, Europhys. Lett. 60(3), 481-487.

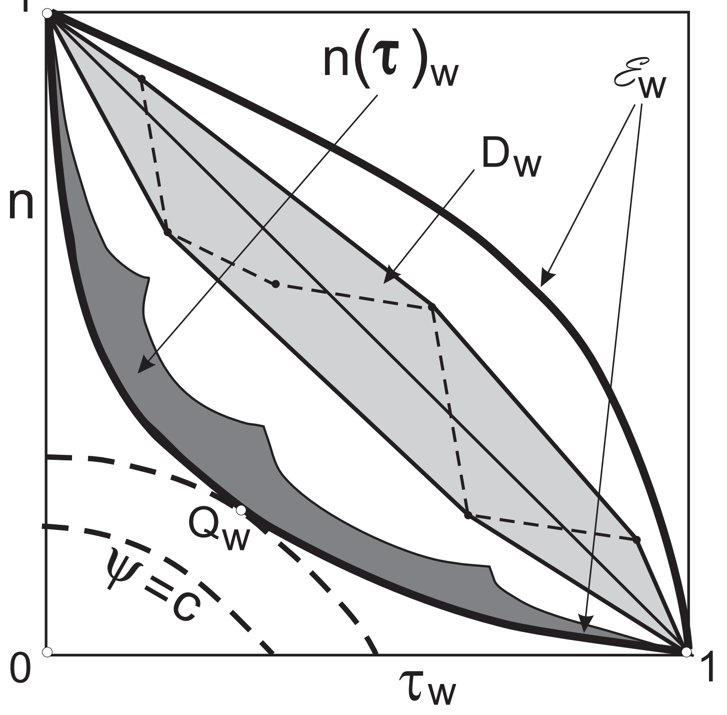
Zechar, J.D. and Jordan, Th.,H. (2008), *Testing alarm-based earthquake predictions*, Geophys. J. Int. 172, 715-724

Figure captions

Fig. 1. Space-time prediction characteristics: n vs. $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_k)$ (the horizontal axis is multidimensional)
Notation: $\mathcal{E}(I)$ represents all strategies based on the data I, the hyperplane D represents the trivial strategies (random guesses), and the surface $n(\boldsymbol{\tau})$ the optimal strategies (the error diagram). The level sets of the loss function $\varphi(n, \boldsymbol{\tau})$ are shown by dashed lines, the characteristic of the optimal prediction is a tangent point Q between $n(\boldsymbol{\tau})$ and the suitable level set of φ.

Fig. 2. The reduced error diagram: n vs. $\tau_w = \sum_{i=1}^{k} \tau_i w_i$
Notation: $\mathcal{E}_w$ contoured by bold lines represents all strategies $\mathcal{E}$ in the (n, τ_w) coordinates; the stippled zone D_w represents the trivial strategies; the broken line within D_w illustrates the method used to construct D_w, see **3.3**; the filled zone is the image of the $n(\boldsymbol{\tau})$ diagram; isolines of the loss function $\varphi = \psi(n, \tau_w)$ are shown by dashed lines; φ yields the optimal characteristics $Q_w = (n, \tau_w)$.


1
n
n(𝛕)
ℰ(I)
D
Q
φ=c*
0
1
𝛕=(τ₁,...τₖ)


1
n(**τ**)$_w$
$\mathscr{E}_w$
D$_w$
n
Q$_w$
$\psi=c$
0
τ_w
1